View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2006 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2006

A Framework for Locating and Analyzing Hate Groups in Blogs

Michael Chau
The University of Hong Kong Pokfulam, mchau@business.hku.hk

Jennifer Xu
Bentley College Waltham, MA, jxu@bentley.edu

Follow this and additional works at: http://aisel.aisnet.org/pacis2006

Recommended Citation
Chau, Michael and Xu, Jennifer, "A Framework for Locating and Analyzing Hate Groups in Blogs" (2006). *PACIS 2006 Proceedings*. 60.
http://aisel.aisnet.org/pacis2006/60

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2006 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

A Framework for Locating and Analyzing Hate Groups in Blogs

Michael Chau
School of Business
The University of Hong Kong Pokfulam, Hong Kong
mchau@business.hku.hk

Jennifer Xu
Dept. of Computer Information Systems Bentley College
Waltham, MA 02452, USA
jxu@bentley.edu

Abstract

As blogs have become one of the fastest growing types of Web-based media, bloggers can express their opinions and emotions more freely and easily than before. In the blogspace, many communities have emerged, which include racists and hate groups that are trying to share their ideology, express their views, or recruit new group members. It is important to analyze these cyber communities, defined based on group membership and subscription linkages, in order to monitor for activities that are potentially harmful to society. While Web mining and network analysis techniques have been widely used to analyze the content and structure of the Web sites of hate groups on the Internet, these techniques have not been applied to the study of hate groups in blogs. In this research, we propose a framework to address this problem. The framework consists of four modules, namely blog spider, information extraction, network analysis, and visualization. We applied this framework to identify and analyze a selected set of 28 anti-Blacks hate groups (820 bloggers) on Xanga, one of the most popular blog hosting sites. Our analysis results revealed some interesting demographical and topological characteristics in these groups, and identified at least two large communities on top of the smaller ones. We suggest that the proposed framework can be generalized and applied to blog analysis in other domains.

Keywords: Web mining, social network analysis, Web communities, hate groups

1. Introduction

Blog is a Web-based publication that allows users to add content easily and periodically, normally in reverse chronological order. Blogs have become increasingly popular in recent years, partly due to the availability of easy-to-use blogging tools and free blog hosting sites, such as www.blogger.com, www.xanga.com, and www.livejournal.com. These tools also support the linking to other pages or the posting of comments to blogs (Blood, 2004). Instead of having a few people being in control of the discussion (like in traditional Internet forums), blogs basically allow anyone to express their ideas and thoughts freely in one's own blog space.

There are many communities in the blogspace. These could be support communities such as those for technical support or educational support (Nardi *et al.*, 2004), or groups of bloggers who already knew each other in other context, such as a group for a high school or a company. In addition, there are also communities formed by people who share common interests or opinions. Many free blog hosting sites have the function to allow bloggers to link to each other to form explicit groups. Similar to other Web-based media such as Web sites, discussion forums, or chat rooms where hate groups are present (Anti-Defamation League, 2001; CNN, 1999; Glaser *et al.*, 2002), there are also hate groups in blogs that are formed by bloggers who are racists or extremists. The consequences of the formation of such groups on the Internet cannot be underestimated. Hate groups or White supremacist groups like the Ku Klux Klan have started to use the Internet to spread their beliefs, recruit new members, or even advocate hate crimes with considerable success (Anti-Defamation League, 2001). The Web has allowed these groups to reach much further into society than ever before. Young people, the major group of bloggers, are more likely to

be affected and even "brainwashed" by ideas propagated through the Web as a global medium. Hatred and extremism ideas could easily be embedded into their minds to make them become members of these hate groups or even conduct hate crimes.

To investigate the cyber activities of hate groups in blogs, it is important to devise an efficient and effective way to identify these groups, extract the information of their members, and explore their relationships. In recent years, advanced techniques such as text mining, Web mining, and social network analysis have been widely used for cyber crime and terrorism analysis in recent years (e.g., Chen *et al*. 2004). For example, network analysis techniques have been used to study the relationships between extremist Web sites on the Internet (Zhou *et al*., 2005). However, the application of these techniques to blog analysis on the Web is a new area and no prior research has been published. We suggest that these techniques can be used to help identify and analyze hate groups on the Web, particularly in blogs. In this paper, we propose a semi-automated framework that combines blog spidering and social network analysis techniques to facilitate such analysis.

The rest of the paper is organized as follows. In Section 2 we review the research background of hate group analysis and related research in text mining, Web mining, and social network analysis. We pose our research questions in Section 3, and a semi-automated framework for hate group analysis in blogs is presented in Section 4. In Section 5 we present a case study that we have performed on a popular blog hosting site based on the proposed framework and discuss our analysis findings. Lastly, we conclude our research in Section 6 and suggest some directions for future work.

2. Research Background

In this section we will review the background of on the Web and in blogs. We also review relevant techniques in Web mining and social network analysis that have been applied in analyzing Web contents.

2.1 Cyberhate and Blogs

Hate crimes have been one of the long-standing problems in the United States because of various historical, cultural, and political reasons. It has been reported that 60% of hate criminals are youths (Levin & McDevitt, 1993), who are, perhaps unfortunately, also one of the largest groups of Internet users. Hate groups have been increasingly using the Internet to express their ideas, spread their beliefs, and recruit new members (Lee & Leets, 2002). Glaser *et al*. (2002) suggest that racists often express their views more freely on the Internet. The Hate Directory (Franklin, 2005) compiles a list of hundreds of Web sites, files archives, newsgroups, and other Internet resources related to hate and racism. Several studies have investigated Web sites that are related to racism or White supremacy (e.g., Lee & Leets (2002); Gerstenfeld *et al*. (2003)). Burris *et al.* (2000) systematically analyzed the networks of Web sites maintained by white supremacist groups and found that this network had a decentralized structure. Zhou *et al*. (2005) used software to automate the analysis of the content of hate group Web sites and the linkage among them. They found that one of the major objectives of these Web sites was to share ideology. Online communities such as White Supremacists and Neo-Nazis were identified among these sites.

In recent years, hate groups have emerged in blogs, where highly-narrative messages are popular. Blogs, also known as weblogs, have become increasingly popular in the past few years. In the early days, blogs were used mainly to designate pages where links to other useful resources were periodically "logged" and posted. At that time blogs were mostly maintained by hand (Blood, 2004). After easy-to-use blogging software became widely available in the early 2000's, the nature of blogs has changed and many blogs are more like personal Web sites that contain various types of content (not limited to links) posted in reverse-chronological order. Bloggers often make a record of their lives and express their opinions, feelings, and emotions through writing blogs (Nardi *et al*., 2004). Many bloggers consider blogging as an outlet for their thoughts and emotions. Besides

personal blogs, there are also blogs created by companies. For example, ice.com, an online jewelry seller, has launched three blogs and reported that thousands of people linked to their Web site from these blogs (Hof, 2005).

One of the most important features in blogs is the ability for any reader to write a comment on a blog entry. On most blog hosting sites, it is very easy to write a comment, in a way quite similar to replying to a previous message in traditional discussion forums. The ability to comment on blogs has facilitated the interaction between bloggers and their readers. On some controversial issues, like those related to racism, it is not uncommon to find a blog entry with thousands of comments where people disputing back and forth on the matter.

Cyber communities have also been formed in blogs. Communities in blogs can be categorized as explicit communities or implicit communities, like other cyber communities on the Web (Kumar *et al*., 1999). Explicit communities in blogs are the groups, or blogrings, that bloggers have explicitly formed and joined. Most blog hosting sites allow bloggers to form a new group or join any existing groups. On the other hand, implicit communities are not explicitly defined as groups or blogrings by bloggers. Instead, these communities are identified by the interactions among bloggers, such as subscription, linking, or commenting. For example, a blogger may subscribe to another blog, meaning that the subscriber can get updates when the subscribed blog has been updated. A blogger can also post a link or add a comment to another blog, which are perhaps the most traditional activities among bloggers. These interactions signify some kind of connection between two bloggers. Similar to the analysis of hyperlinks among Web pages to identify communities (Chau *et al*., 2005; Chau *et al*., forthcoming), analysis of these types of connection between bloggers could also identify cyber communities and their relationships.

2.2 Web Mining and Social Network Analysis

In recent years, Web mining techniques have been widely adopted from data mining, text mining, and information retrieval research and applied to various Web applications. Web mining research can be classified into three categories: Web content mining, Web structure mining, and Web usage mining (Kosala & Blockeel, 2000; Chen & Chau, 2004). Web content mining refers to the discovery of useful information from Web contents, including text, images, audio, video, etc. Web content mining research includes resource discovery from the Web, document categorization and clustering, and information extraction from Web pages. Web structure mining studies the model underlying the link structures of the Web. It usually involves the analysis of in-links and out-links information of a Web page, and has been used for search engine result ranking and other Web applications. Google's PageRank (Brin & Page, 1998) and HITS (Kleinberg, 1998) are the two most widely used algorithms. Web usage mining focuses on using data mining techniques to analyze search logs or other activity logs to find interesting patterns. One of the main applications of Web usage mining is its use to learn user profiles.

In Web mining research, Web spiders have been widely used to traverse the Web and collect Web pages for further analysis. Spiders, also known as crawlers, wanderers, or Webbots, have been defined as "software programs that traverse the World Wide Web information space by following hypertext links and retrieving Web documents by standard HTTP protocol" (Cheong, 1996). Since the early days of the Web, spiders have been widely used to build the underlying databases of search engines, to perform personal search, to archive particular Web sites or even the whole Web, or to collect Web statistics. Chau and Chen (2003) provide a review of Web spider research.

Various methods have been proposed in Web structure mining research to identify Web communities (Gibson *et al.*, 1998; Kumar *et al.*, 1999). Many of these methods are rooted in the HITS algorithm (Kleinberg, 1998). Kumar *et al.*

(1999) propose a trawling approach to find a set of core pages containing both authoritative and hub pages for a specific topic. The core is a directed bipartite subgraph whose node set is divided into two sets with all hub pages in one set and authoritative pages in the other. The core and the other related pages constitute a Web community (Gibson *et al.*, 1998). Treating the Web as a large graph, the problem of community identification can also be formulated as a minimum-cut problem, which finds clusters of roughly equal sizes while minimizing the number of links between clusters (Flake *et al.*, 2000; Flake *et al.*, 2002). Realizing that the minimum-cut problem is equivalent to the maximum-flow problem, Flake *et al.* (2000) formulate the Web community identification problem as an *s-t* maximum flow problem, which can be solved using efficient polynomial time methods. Hierarchical clustering methods have also be proposed to partition networks, especially unweighted networks such as the Web in which hyperlinks do not have associated weights (e.g., Radicchi *et al.*, 2004).

On the other hand, a recent movement in statistical analysis of network topology (Albert & Barabási, 2002) has brought new insights and research methodology to the study of network structure. Networks, regardless of their contents, are classified into three categories: *random network* (Bollobás, 1985), *small-world network* (Watts & Strogatz, 1998), and *scale-free network* (Barabási & Alert, 1999). In a random network the probability that two randomly selected nodes are connected is a constant *p*. As a result, each node has roughly the same number of links and nodes are rather homogenous. In addition, communities are not likely to exist in random networks. Small-world networks, in contrast, have a significantly high tendency to form groups and communities. Most empirical networks including social networks, biological networks, and the Web have been found to be nonrandom networks. In addition, many of these networks are also scale-free networks (Barabási & Alert, 1999), in which a large percentage of nodes have just a few links, while a small percentage of the nodes have a large number of links. Thus, nodes in scale-free networks are not homogenous in terms of their links, and some nodes become hubs or leaders that play important roles in the functioning of the network. The Web has been found to have both

small-world and scale-free properties (Albert & Barabási, 2002). Researchers have been employing social network analysis methods to analyze the structure of the Web (Kumar *et al.*, 2002).

Social network analysis (SNA) is a sociological methodology for analyzing patterns of relationships and interactions between social actors in order to discover the underlying social structure (Wasserman & Faust, 1994). Not only the attributes of social actors, such as their age, gender, socioeconomic status, and education, but also the properties of relationships between social actors, such as the nature, intensity, and frequency of the relationships, are believed to have important implications to the social structure. SNA methods have been employed to study organizational behavior, inter-organizational relations, citation patterns, computer mediated communication, and many other areas.

Recently, SNA has also been used in the intelligence and security domain to analyze criminal and terrorist networks (Krebs, 2001; Xu & Chen, 2004; 2005). When used to mine a network, SNA can help reveal the structural patterns such as the central nodes which act as hubs, leaders, or gatekeepers, the densely-knit communities or groups in which nodes have frequent interactions with each other, and the patterns of interactions between the communities and groups. These patterns often have important implications to the functioning of the network. For example, the central nodes often play a key role by issuing commands or bridging different communities. The removal of central nodes can effectively disrupt a network than peripheral nodes (Albert *et al.*, 2000).

3 Research Questions

We believe that it is an important and timely issue to identify the hate groups in blogs and analyze their relationships. Web mining and network analysis techniques have been used to analyze Web content such as Web pages and hyperlinks; however, we have not been able to identify any prior research in this aspect in the literature. Based on our review, we pose the following research questions: (1) Can we use a semi-automatic framework to identify hate groups in

blogs? (2) What is the pattern of interaction between bloggers in hate groups? (3) Can social network analysis techniques be used to analyze the groups and their relationships in blogs?

4 Proposed Framework

In this section, we propose a semi-automated framework for identifying groups and analyzing their relationships in blogs. The framework is diagrammed in Figure 1. Our framework consists of four main modules, namely Blog Spider, Information Extraction, Network Analysis, and Visualization. The Blog Spider module downloads blog pages from the Web. These pages are then processed by the Information Extraction module. Data about these blogs and their relationships are extracted and passed to the Network Analysis module for further analysis. Finally the Visualization module presents the analysis results to users in a graphical display. In the following, we describe each module in more detail.

4.1 Blog Spider

A blog spider program is first needed to download the relevant pages from the blogs of interest. Similar to general Web fetching, the spider can connect to blog hosting sites using standard HTTP protocol. After a blogring description page or a blog page is fetched, URLs are extracted and stored into a queue. However, instead of following all extracted links, the blog spider should only follow links that are of interest, e.g., links to a group's members, other bloggers, comment links, and so on. Links to other external resources are often less useful in blog analysis. In addition, the spider can use RSS and get notification when the blog is updated. However, this is only necessary when monitoring or incremental analysis is desired.

4.2 Information Extraction

After the blog pages have been downloaded, information extraction can be performed. This includes information related to the blog or the blogger, such as user profiles and date of creation. This can also include linkage information between two bloggers, such as linkage, commenting, or subscription. Because different blogs may have different formats, it is not a trivial task to extract the user profiles, links, comments, and other useful information from blogs. Even blogs hosted on the same hosting site could have considerably different formats as they can be easily customized by each blogger. Pattern matching or entity extraction techniques can be applied. The information extracted can be stored into database for further analysis. Links extracted can be passed back to the blog spider for further fetching.

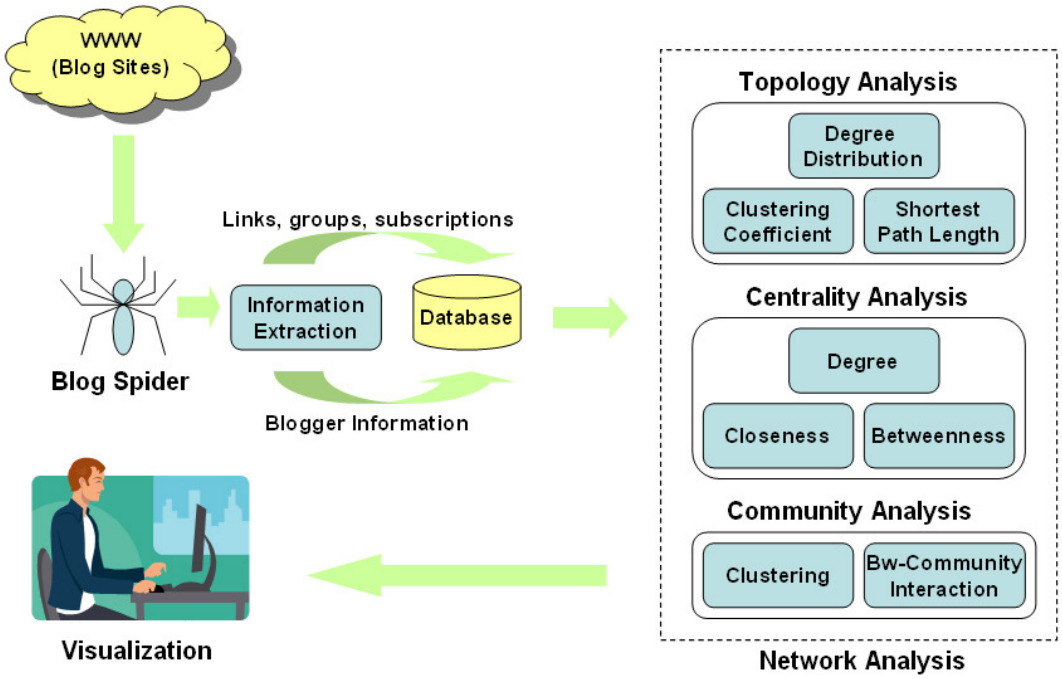

Figure 1. The proposed semi-automated framework for blog link analysis

4.3 Network Analysis

Network analysis is a major component in our framework. In this module we propose three types of analysis: *topological analysis*, *centrality analysis*, and *community analysis*.

The goal of topological analysis is to ensure that the network extracted based on links between bloggers is not random and it is meaningful to perform the centrality and community analysis. We use three statistics that are widely used in topological studies to categorize the extracted network (Albert & Barabási, 2002): *average shortest path length*, *clustering coefficient*, and *degree distribution*. Average path length is the mean of the all-pair shortest paths in a network. It measures the efficiency of communication between nodes in a network. Clustering coefficient indicates how likely nodes in a network form groups or communities. The degree distribution, $P(k)$, is the probability that a node has exactly k links. Another measure related to average path length is the network's global efficiency, which is defined as the average of the inverses of shortest path lengths over all pairs of nodes in a network (Crucitti *et al.*, 2003).

Centrality analysis follows the topological analysis if the extracted network is shown to be a nonrandom network in which node degrees may vary greatly. The goal of centrality analysis is to identify the key nodes in a network. Three traditional centrality measures can be used: *degree*, *betweenness*, and *closeness* (Freeman, 1979). Degree measures how active a particular node is. It is defined as the number of direct links a node has. "Popular" nodes with high degree scores are the leaders, experts, or hubs in a network. In the intelligence and security context, the removal of these key nodes in a criminal or terrorist network is often an effective disruptive strategy (Sparrow, 1991). Betweenness measures the extent to which a particular node lies between other nodes in a network. The betweenness of a node is defined as the number of geodesics (shortest paths between two nodes) passing through it. Nodes with high betweenness scores often serve as gatekeepers and brokers between different

communities. They are important communication channels through which information, goods, and other resources are transmitted or exchanged (Wasserman & Faust, 1994). Closeness is the sum of the length of geodesics between a particular node and all the other nodes in a network. A node with low closeness may find it very difficult to communicate with other nodes in the network. Such nodes are thus more "peripheral" and can become outliers in the network (Sparrow, 1991; Xu & Chen, 2005).

Community analysis is to identify social groups in a network. In SNA a subset of nodes is considered a community or a social group if nodes in this group have stronger or denser links with nodes within the group than with nodes outside of the group (Wasserman & Faust, 1994). A weighted network in which each link has an associated weight can be partitioned into groups by maximizing the within-group link weights while minimizing between-group link weights. An unweighted network can be partitioned into groups by maximizing within-group link density while minimizing between-group link density. In this case, groups are densely-knit subsets of the network. Note that community and groups here do not refer to the explicit groups (blogrings). They refer to a subset of nodes which form implicit clusters through various relationships, even if these nodes do not belong to the same explicit group.

After a network is partitioned into groups, the between-group relationships become composites of links between individual nodes. In SNA, a method called *blockmodeling* is often used to reveal the patterns of interactions between groups (White *et al.*, 1976). Given groups in a network, blockmodel analysis determines the presence or absence of a relationship between two groups based on the link density (Wasserman & Faust, 1994). When the density of the links between the two groups is greater than a predefined threshold value, a between-group relationship is present, indicating that the two groups interact with each other constantly and thus have a strong relationship. By this means, blockmodeling summarizes individual relational details into relationships between groups so that the overall structure of the network becomes more prominent.

4.4 Visualization

The extracted network and analysis results can be visualized using various types of network layout methods. Two examples are *multidimensional scaling* (MDS) and *graph layout* approaches. MDS is the most commonly used method for social network visualization (Freeman, 2000). It is a statistical method that projects higher-dimensional data onto a lower-dimensional display. It seeks to provide a visual representation of proximities (dissimilarities) among nodes so that nodes that are more similar to each other are closer on the display and nodes that are less similar to each other are further apart (Kruskal & Wish, 1978). Various graph layout algorithms, such as the force-directed method, have been developed particularly for drawing aesthetically pleasing network presentations (Fruchterman & Reingold, 1991).

5 A Case Study on Xanga

5.1 Focus and Methods

We applied our framework to perform a case study of hate groups in blogs. We chose to study the hate groups against Blacks because of two reasons. First, the nature of hate groups and hate crimes is often dependent on the target "hated" group. By focusing on a type of hate groups it is possible to identify relationships that are more prominent. Second, among different hate crimes, anti-Black hate crimes have been one of the most widely studied. This allows us to compare our results with previous research in the literature.

We limit our study to blogs on Xanga (www.xanga.com). According to Alexa (2005), Xanga is the second most popular blog host, only after Blogger (www.blogger.com). It is also ranked 17th in traffic (visit popularity) among all Web sites in English. We chose Xanga over Blogger because Xanga has more prominent features to support subscriptions and groups (as blogrings). These features are useful for the identification of hate groups in the blogs and the

relationships between bloggers. Also, it has been suggested that apparently more hate blogs exist in Xanga than in Blogger (Franklin, 2005).

After choosing our focus, we had to identify a set of hate groups in Xanga. We used the search feature in Xanga to semi-automate the task. First, a set of terms related to Black-hatred, such as "KKK", "niggers", "white pride", were identified. We used these terms to search for groups (blogrings) in Xanga that have any of these words in their group name or description. We then checked these groups and filtered out those not related to anti-Black. Groups with only one single member, which were often formed by one blogger but no one else had joined afterwards, were also removed from our list. This resulted in a set of 40 groups. While most of these groups showed some beliefs of racism or White supremacies, we tried to further narrow these down to groups that demonstrated explicit hatred, so as to make sure that our analysis focused on "hate groups". So, we then manually checked these groups and only included those that explicitly mentioned hatred (e.g., "I hate black people", "hate the black race") or used offensive languages (e.g., "nigger beaters", four-letter words) towards the Blacks in their group name or description. Finally we had a list of 28 groups.

Spiders were used to automatically download the description page and member list of each of these groups. A total of 820 bloggers were identified from these 28 groups. The spiders then further downloaded the blogs of each of these bloggers. The extraction program was then executed to extract the information from each blogger, including user id, real name, date of creation, date of birth, city, state, and country. One should note that these data were self-reported; they could be fraud or even missing.

The extraction program also analyzed the relationship between these bloggers. In this study, two types of relationships were extracted:

1. Group co-membership: two bloggers belong to the same group (blogring). This is an undirected relationship with an integer weight (based on the number of groups shared by the two bloggers). As using all co-membership links would result in a very large network that corresponds to

the original blogring information, we only included links where the weight is at least 2.
2. Subscription: blogger A subscribes to blogger B. This is a directed, binary relationship.

5.2 Analysis

After collecting the blogs and extracting information from them, we performed demographical, topological analysis, and social network analysis on the data set in order to reveal the characteristics of these groups and study whether any patterns exist. Visualization was also applied to present the results. We discuss the details of our analysis in the following.

5.2.1 Demographical Analysis

We provide a brief summary of the demographical information of the bloggers of interest and the growth patterns of the blog space of hate groups. As in many other Internet-based media such as forums and chat rooms, the real identities of bloggers are unknown. Thus, the self-reported demographical information of bloggers is also subject to the problems of anonymity. However, since blogs are often personal online diaries many bloggers still choose to release partial information about their demographics such as gender and country. Six hundred and fifty-nine bloggers out of the 820 in our data set have explicitly indicated their gender. Sixty three percent of them are male and 37% female. These bloggers are from various countries, with 81.9% from the United States and the remaining18.1% from 45 different countries. It can be seen that hate groups are dominated by male bloggers from the United States. However, one should note that this finding is based on the problematic source of demographic information.

We also analyzed the growth of the hate group blogs over the years. Unlike demographical information, the exact time when a blogger registered on the blog hosting site (Xanga) is recorded by the server and thus is generally not subject to fraud. We found that the number of “hate” bloggers increased steadily between 2003 and the third quarter of 2004. The number has fluctuated since the forth quarter of 2004. This may be because some bloggers who have recently registered have not joined those popular blogrings or have not formed into large

communities. As a result, some of them were not included in the data set after we filtered the raw data. The finding suggests that hate groups have been gaining popularity in blog space over years. Such a trend should not be underestimated because the ideas, beliefs, and opinions advocated by racists and extremists may pose potential threats to the society.

5.2.2 Topological Analysis

When analyzing the topology of the network, we ignored the weight of co-membership relationship and the direction of subscription. We connected two nodes if they belonged to at least two common groups or one subscribed to the other. So, there could only be at most one link between a pair of nodes. The resulting network was an unweighted, undirected network consisting of 1193 links. This network was not a connected graph in that it consisted of several disjoint components, between which no link existed. The largest connected component, often called a giant component in graph theory (Bollobás, 1985), contained 273 nodes connected by 1115 links. This giant component was a rather dense graph with an average node degree of 8.2.

	Average Shortest Path Length	Global Efficiency	Clustering Coefficient
Giant Component	3.62	0.33	0.37
Random Counterpart	2.89 (0.03)	0.37 (0.00)	0.03 (0.00)

Table 1. The topological properties of the giant component (number in the parentheses are standard deviations).

We performed topological analysis for the giant component. Table 1 provides the statistics of the average shortest path length, global efficiency, and clustering coefficient. To compare the giant component with its random graph counterpart, we generated 30 random networks consisting of the same number of nodes (273) and links (1115) with the giant component. The resulting statistics are also reported in Table 1. It shows that the giant component is less efficient than its random graph counterpart. On average, a node in the giant component must take 0.75 more steps than in the random graph to reach another arbitrary node. Accordingly, the global efficiency of the giant component is also relatively low.

However, the giant component has a significantly higher clustering coefficient, which is 31 times more than its random graph counterpart. This implies that the giant component is a small-world, in which densely-knit communities are very likely to exist.

The degree distribution of the giant component also seems to follow a power-law distribution (see Figure 2). The most distinctive feature of a power-law distribution curve is its long tail for large degree (*k*), which is significant different from a bell-shaped Poisson distribution. The long tail indicates that a small number of nodes in the network have a large number of links and they are the key nodes to identify.

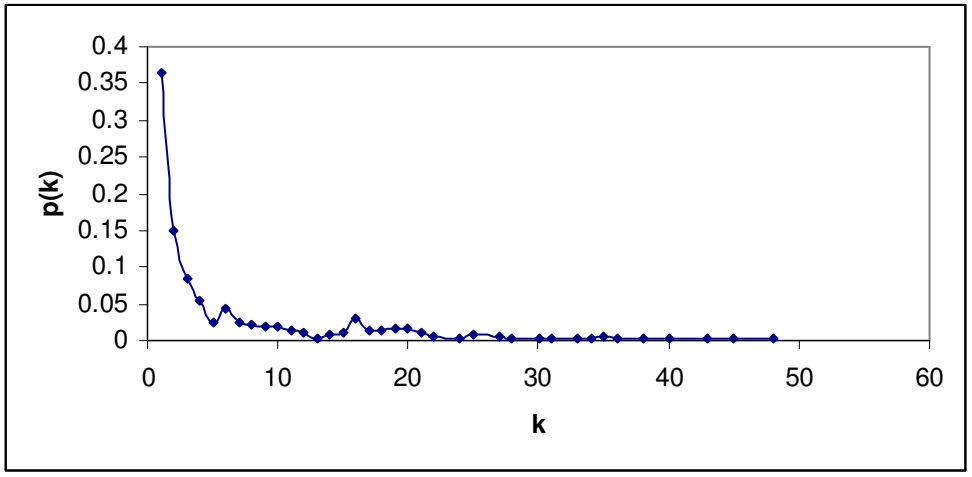

Figure 2. The degree distribution of the giant component

5.2.3 SNA and Visualization

We performed centrality and community analysis using a prototype system we developed (Xu & Chen, 2005). The system has a graphical user interface to facilitate interaction between users and the system (see Figure 3). The user interface visualizes the network and presents the results of social network analysis. In the visualization, each node represents a blogger. A straight line connecting two nodes indicates that either the two corresponding are co-members in more than one blogring, or one blogger subscribes to the other.

The layout of the network is determined using the MDS method. In order to position nodes which are likely to belong to the same community close to each

other on the display, we assigned each link an "edge clustering coefficient", which measures the likelihood of two incident nodes of the link to form a cluster (Radicchi *et al*., 2004).

The community analysis can be performed by adjusting the "level of abstraction" slider at the bottom of the panel. At the lowest level of abstraction, each individual node and link are presented. As the abstraction level increases, the system employs hierarchical clustering method to gradually merge nodes, which are connected by links of high edge clustering coefficients. When the highest abstraction level is reached, the whole giant component becomes a big community.

At any level of abstraction, a circle represents a community. The size of the circle is proportional to the number of bloggers in the community. Straight lines connecting circles represent between-group relationships, which are extracted using blockmodel analysis. The thickness of a line is proportional to the density of the links between the two corresponding communities.

Figure 3 presents the giant component at its lowest abstraction level. The bloggers who have the highest degree and betweenness scores are highlighted and labeled with their usernames. These bloggers are those who may participate in multiple blogrings or have many subscription relationships with other bloggers. It is interesting to see from this figure that in addition to joining explicit groups (blogrings), bloggers have also formed implicit communities through co-membership and subscription. Three circles of nodes are apparently communities in which bloggers share many common memberships.

Because the communities can be analyzed at different levels, the two big communities may consist of smaller communities. For example, the inner structure of the bigger community shows that two smaller circles of nodes, which may also be communities, exist in the community. A blogger who ranks the highest in degree in the community would likely be the leader of the community. Also, bloggers with a high in-degree would be those who can disseminate their

ideas and opinions through their blogs easily to many subscribers. These "leader" bloggers are the ones who should be substantially monitored by authorities which hope to regulate the activities of hate groups in blogs.

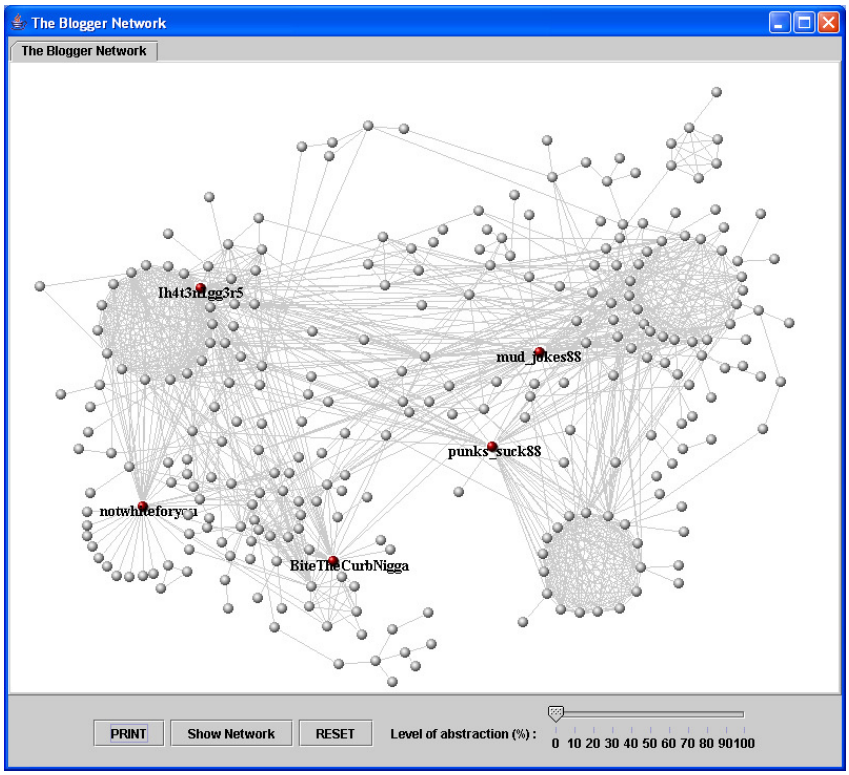

Figure 3. The prototype system for SNA and visualization, showing the giant component with both co-membership and subscription links. The highlighted nodes are those who have large degree or betweenness scores.

6 Conclusion and Future Directions

In this paper, we have discussed the problems of the emergence of hate groups and racism in blogs. We have proposed a semi-automated framework for blog analysis, and applied this framework to investigate the characteristic and structural relationships among the hate groups in blogs. We believe that this research is timely and important to the security of society. By disseminating both explicit and implicit hatred messages through blogs, racists can easily target youths with a ubiquitous coverage – basically anyone who has access to the Internet – that was never possible in the past. Youths are often easily influenced by these messages and could eventually become terrorists and pose a threat to

our society (Blazak, 2001). Our study has provided a framework that could facilitate the analysis of law enforcement and social workers in studying and monitoring such activities.

While the framework has been proposed and studied in the context of hate group analysis, we have tried to keep the framework general such that it is not specific to such analysis. We believe that the framework can also be applied to other content and network analysis research that involves blog mining, which we believe would be an increasingly important field for various applications. These applications include not only other security informatics research but also applications in other domains such as marketing analysis and business intelligence analysis (Chau *et al.*, 2005).

We are extending our study in two major directions. First, the current study only investigated the hate group activities on one single blog site, Xanga. Although Xanga has been reported to have the most number of blogs associated with hate groups (Franklin, 2005), a further study that includes other popular sites such as Blogger would be more comprehensive. Second, only two types of relationships, namely co-membership and subscription, were considered in the present study. It would be interesting to expand our study to include other types of relationships, such as commenting and hyperlinking, in the network analysis. Inclusion of these relationships could reveal other implicit linkages among the bloggers.

Acknowledgement

This research has been supported in part by a grant from the University of Hong Kong Seed Funding for Basic Research, "Searching and Analyzing Blogs for Competitive Advantages," 10206775 (PI: M. Chau), January 2006 – December 2007. We would like to thank Porsche Lam and Boby Shiu of the University of Hong Kong for their participation and support in this project.

References

Albert, R. & Barabási, A.-L. (2002). "Statistical Mechanics of Complex Networks." *Reviews of Modern Physics*, 74(1), 47-97.

Albert, R., Jeong, H., Barabási, A.-L. (2000). "Error and Attack Tolerance of Complex Networks." *Nature*, 406, 378-382.

Alexa (2005). "Top English Language Sites," [Online] Retrieved from http://www.alexa.com/site/ds/top_sites?ts_mode=lang&lang=en on October 7, 2005.

Anti-Defamation League (2001). "Poisoning the Web: Hatred Online," [Online] Retrieved from http://www.adl.org/poisoning_web/poisoning_toc.asp on October 7, 2005.

Barabási, A.-L., Albert, R., & Jeong, H. (1999). "Mean-Field Theory for Scale-Free Random Networks." *Physica A*, 272, 173-187.

Blazak, R. (2001). "White Boys to Terrorist Men: Target Recruitment of Nazi Skinheads," *American Behavioral Scientist*, 44(6), 982-1000.

Blood, R. (2004). "How Blogging Software Reshapes the Online Community," *Communications of the ACM*, December 2004, 47(12), 53-55.

Bollobás, B. (1985). *Random Graphs*. London, Academic.

Brin, S. & Page, L. (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Proceedings of the 7th WWW Conference*, Brisbane, Australia, April 1998.

Burris, V., Smith, E., & Strahm, A. (2000). "White Supremacist Networks on the Internet," *Sociological Focus*, 33(2), 215-235.

Chau, M. & Chen, H. (2003). "Personalized and Focused Web Spiders," in N. Zhong, J. Liu, & Y. Yao (Eds), *Web Intelligence*, Springer-Verlag, 197-217.

Chau, M., Shiu, B., Chan, I., & Chen, H. (2005). "Automated Identification of Web Communities for Business Intelligence Analysis," in *Proceedings of the Fourth Workshop on E-Business (WEB 2005)*, Las Vegas, USA, December, 2005.

Chau, M., Shiu, B., Chan, I., & Chen, H. (forthcoming). "Redips: Backlink Search and Analysis on the Web for Business Intelligence," *Journal of the American Society for Information Science and Technology*, accepted for publication, forthcoming.

Chen, H. and Chau, M. (2004). "Web Mining: Machine Learning for Web Applications," *Annual Review of Information Science and Technology*, 38, 289-329, 2004.

Chen, H., Chung, W., Xu. J., Wang, G., Qin, Y., & Chau, M. (2004). "Crime Data Mining: A General Framework and Some Examples," *IEEE Computer*, 37(4), 50-56.

Cheong, F. C. (1996). *Internet Agents: Spiders, Wanderers, Brokers, and Bots*. Indianapolis, Indiana, USA: New Riders Publishing.

CNN (1999). "Hate Group Web Sites on the Rise," *CNN News* [Online] Retrieved from http://edition.cnn.com/US/9902/23/hate.group.report/index.html on October 7, 2005.

Crucitti, P., Latora, V., Marchiori, M., & Rapisarda A. (2003). "Efficiency of Scale-Free Networks: Error and Attack Tolerance." *Physica A*, 320, 622-642.

Flake, G. W., Lawrence, S., & Giles, C. L. (2000). "Efficient Identification of Web Communities." In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining* (ACM SIGKDD 2000), Boston, MA.

Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2002). "Self-Organization and Identification of Web Communities." *IEEE Computer,* 35(3), 66-71.

Franklin, R. A. (2005). "The Hate Directory" [Online] Retrieved from http://www.bcpl.net/~rfrankli/hatedir.htm on October 7, 2005.

Freeman, L. C. (1979). "Centrality in Social Networks: Conceptual Clarification." *Social Networks*, 1, 215-240.

Freeman, L. C. (2000). "Visualizing Social Networks." *Journal of Social Structure*, 1(1).

Fruchterman, T. M. J. & Reingold, E. M. (1991). "Graph Drawing by Force-Directed Placement." *Software-Practice & Experience*, 21(11), 1129-1164.

Gerstenfeld, P. B., Grant, D. R., & Chiang, C. P. (2003). "Hate Online: A Content Analysis of Extremist Internet Sites," *Analyses of Social Issues and Public Policy*, 3, 29-44.

Gibson, D., J. Kleinberg, & Raghavan, P. (1998). "Inferring Web Communities from Link Topology." In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, Pittsburgh, PA.

Girvan, M. & Newman, M. E. J. (2002). "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Science of the United States of America*, 99, 7821-7826.

Glaser, J., Dixit, J., & Green, D. P. (2002). "Studying Hate Crime with the Internet: What Makes Racists Advocate Racial Violence?" *Journal of Social Issues*, 58(1), 177-193.

Hof, R. (2005). "Blogs on Ice: Signs of a Business Model?" *Business Week Online – The Tech Beat*, June 2, 2005. [Online] Retrieved from http://www.businessweek.com/the_thread/techbeat/archives/2005/06/blogs_on_ice_si.html on October 7, 2005.

Kleinberg, J. (1998). "Authoritative Sources in a Hyperlinked Environment," in *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, San Francisco, California, USA, Jan 1998, pp. 668-677.

Kosala, R. & Blockeel, H. (2000). "Web Mining Research: A Survey," *ACM SIGKDD Explorations*, 2(1), 1-15.

Krebs, V. E. (2001). "Mapping Networks of Terrorist Cells." *Connections*, 24(3), 43-52.

Krupka, G. R. & Hausman, K. (1998). "IsoQuest Inc.: Description of the NetOwlTM extractor system as used for MUC-7," in *Proceedings of the Seventh Message Understanding Conference*, April 1998.

Kruskal, J. B. & Wish, M. (1978). *Multidimensional Scaling*. Beverly Hills, CA, Sage Publications.

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). "Trawling the Web for Emerging Cyber-Communities." *Computer Networks*, 31(11-16), 1481-1493.

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (2002). "The Web and Social Networks." *IEEE Computer*, 35(11), 32-36.

Lee, E., & Leets, L. (2002). "Persuasive Storytelling by Hate Groups Online: Examining Its Effects on Adolescents," *American Behavioral Scientist*, 45, 927-957.

Levin, J., & McDevitt, J. (1993). *Hate crimes: The Rising Tide of Bigotry and Bloodshed*. New York: Plenum.

Nardi, B. A., Schiano, D. J., Gumbrecht, M., & Swartz, L. (2004). "Why We Blog," *Communications of the ACM*, December 2004, 47(12), 41-46

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). "Defining and Identifying Communities in Networks." *Proceedings of the National Academy of Science of the United States of America*, 101, 2658-2663.

Sparrow, M. K. (1991). "The Application of Network Analysis to Criminal Intelligence: An Assessment of the Prospects." *Social Networks*, 13, 251-274.

Wasserman, S. & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge, Cambridge University Press.

Watts, D. J. & Strogatz, S. H. (1998). "Collective Dynamics of 'Small-World' Networks." *Nature*, 393, 440-442.

White, H. C., Boorman, S. A., & Breiger, R. L. (1976). "Social Structure from Multiple Networks: I. Blockmodels of Roles and Positions." *American Journal of Sociology*, 81, 730-780.

Xu, J. J. & Chen, H. (2004). "Fighting Organized Crime: Using Shortest-Path Algorithms to Identify Associations in Criminal Networks." *Decision Support Systems*, 38(3), 473-487.

Xu, J. J. & Chen, H. (2005). "CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery." *ACM Transactions on Information Systems*, 23(2), 201-226.

Zhou, Y., Reid, E., Qin, J., Chen, H., & Lai, G. (2005). "US Domestic Extremist Groups on the Web: Link and Content Analysis," *IEEE Intelligent Systems*, 20(5), 44-51.